

开放数据的集成应用研究*

顾立平
(中国科学院国家科学图书馆　北京 100190)

【摘要】在系统性梳理开放数据案例的基础上,说明开放数据发展中面临的科技信息政策问题,提出图书馆学的信息组织原则和方法是面对未来开放数据管理的一项重要预科学基础。

【关键词】语义网　资源描述框架　关联开放数据　用户配置文件　科技创新　开放治理　DBpedia　SPARQL
【分类号】G250

Research on the Applications and Integration of Open Data

Ku Liping
(National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

【Abstract】Based on the systematically reviewed cases of the open data, the paper describes the science and technology policy issues as the bottlenecks of the open data development, and indicates that the principle and method of the information organization of the library science as an important basic of the pre-science for the open data management in future.

【Keywords】Semantic Web　RDF　Linked open data　Profile　Science and technology innovation　Opening govern　DBpedia　SPARQL

1　科技创新与知识社会的发展趋势

越来越多人认为社会的模型正在被改变,从工业模型到知识模型,而知识模型最主要的改变在于原料质变,也就是数据和信息交流方式的改变[1]。近年来随着开放获取运动(Open Access, OA)和开放档案倡议(Open Archives Initiative, OAI)的提出和推进,在信息管理领域中,开放数据(Open Data, OD)和关联开放数据(Linked Open Data, LOD)成为重要举措[2]。例如,过去地理勘探和地理研究必须单独依靠文献阅读和专项研究计划收集部分有限的数据,现在通过使用 OAI 以及 LOD 原则和标准建立的地理空间网络应用程序,能解决许多应用学科快速发现地理信息资源的需求,通过复合异构网络资源进行灵活的数据集成[3]。其对于科学研究、社会管理乃至国防边界测量都发挥了巨大作用。

化学界的蓝方尖塔(Blue Obelisk)运动目的是促进化学软件之间的互操作性、鼓励开放源码开发人员之间的合作,发展化学研究社群的资源和开放标准[4]。科研人员特别是善于从化学信息计量中寻找问题解决方案的科研人员,观察到化学原料之外的"化学类信息原料"对于学科发展具有重要影响。尽管开放标准和开源软件还存在许多挑战,但是蓝方尖塔运动已经为化学科研人员汇聚了许多免费使用的有用资源。因此,认识并且准备开放数据的应用与集成,是一项已经发生并且仍在发展的关键信息组织议题,这不是地理、化学或者其他专门学科的特殊现象,而是全体知识经济社会发展的一项主要动力。

收稿日期:2012-06-11
*本文系中国科学院国家科学图书馆指向性人才项目"科技信息政策研究与咨询中心"(项目编号:馆 1203)的研究成果之一。

从科技创新与知识社会的发展趋势来看,开放数据的应用将成为一项推动社会变革的原料。研究开放数据有助于图书情报机构掌握此类技术,并且理解科技创新和国际社会的变化脉动,提早准备相应的服务支援工作。

2 科研需求所产生的科研工具

开放数据有两个层面:上层的数据格式,例如从一个专有格式转化成资源描述框架(Resource Description Framework, RDF)数据;底层的本体学习的关系模式,即如何有效应用语义网(Semantic Web, SW)与关联数据(Linked Data, LD)[5]。目前多数采用释出第一层开放数据(例如:MARC、DC、RDA 等),而对底层学科专业的关系模式采用相对封闭的管理模式,以确保数据品质和应用效果。各个机构或者网站根据实际需求,设计并且制定第二层关联开放数据规则,而释出第一层开放数据彼此共享。当前的科学发展逐渐并且已经朝向数据密集型科学,由于逐渐认识到开放科学的重要性,组建开放数据的社群和运动,相关的科研工具也相继出现。

2.1 全球暖化的海冰监测需求

地球或者陆地消失,人类将不复存在。在全球环境与空间安全监测计划中,欧洲航天局(European Space Agency, ESA)的欧洲雷达观测站,在两个极地轨道卫星延续和改进的发展报告书中,规划了如何进行数据处理、分发和归档等改革制度,其中发展政策包括提供开放数据以服务海洋监测、海冰监测和检测下沉和山体滑坡等紧急反馈措施[6]。全球各地的科学家、信息分析师甚至只是业余爱好地质观测的网民,可以在这个平台上收集数据、建立模型、进行预测。开放数据并不为科研项目和经费支持的团队独享的意义是:任何人都能检验科研成果和运用科研成果,从而加大科研群体的公信力和影响力。哪怕只是非专业人士所提供的实验数据和结果,也可以在开放社群中经过科学验证无误后做出贡献。

2.2 全球板块推移的地震预测

在一个正常的断层地震序列记录过程中,有可能快速收集引发破坏性地震的数据,并且快速反应地震应急情报。欧洲正在建立不同网络之间通信平台的共同网络运营战略,使地震构造和地震物理研究人员在不同的开放数据档案中,更容易地寻找相关的危险性信息[7]。由于过去数据的零散收集,使得参数估计检验的取材有限,然而预测科学主要的逻辑是所有证据指向同一个方向,而且证据链越多越好(预测更加准确快速)。因此,拉开预测时间和发生时间的差距则依赖更多有效数据的完整集成,目前,开放数据能够打破各种隔阂,是取得这种数据的最佳途径。

2.3 全球物种观察和药物开发

人类对抗疾病有数千年历史,目前病毒已经全球流行并且交互繁殖,而人们却尚未建立好共同合作研究的机制,所幸有 ChEMBL 这种开放的数据资料库为先驱者。它包含类似药物的生物活性化合物的大量信息,这些数据来自定期公布的主要文献,然后予以标准化,最大限度地提高数据质量,使化学生物学和药物发现得到广泛应用[9]。根据 IMEX 联盟(The International Molecular Exchange Consortium)订立的分子数据交换原则[10],EBI 是一个建立分子相互作用的开源数据库,到 2011 年 9 月为止,已经囊括了大约 275 万条来自5 000余种出版物上的数据[11],它的开放数据来自文献数据或者直接数据。通过访问这个网站[12],可以取得完整的源代码与开放数据,其重要意义不仅是生物医学领域的开放数据机制,还是开放科学中的合作共享典范。

2.4 跨越文献流通限制建立蛋白结构数据集成

开放数据库中,基因表达数据的数量不断增长,跨越不同的数据集,大规模收集这些基因表达的相似之处,产生新的关系,可能有助于观察特定模式的原因和后果的假设建立[8]。根据这个逻辑,生物医学研究目前正在积极地从事开放数据运动。例如剑桥大学的 CrystalEye 项目,从网络资源中自动生成关于晶体结构数据的结构化 XML 的开放数据,提供浏览、搜索和新知快报[13]。可以说,在开放科学中,专家的定义不在于拥有多少头衔、掌握多少别人没有的资源,而在于使用少量资源(或者日益增长的开放数据)做出贡献。

2.5 跨越语种限制建立大规模中文维基资源

语义网的目标是建立"数据网络"使得机器了解网络上的信息;关联开放数据(LOD)项目鼓励个人和组织在网络上发布各种公开的数据集,再通过资源描述框架(RDF)促进语义网的快速发展。然而,根据字母拼音系统设计的 DBpedia 并未考虑到中文这种非西

方的语言结构，从而阻碍了跨语言资源的知识共享。浙江大学的团队为了解决这个问题，建立了一个基于 Wiki 类别和 InfoBoxes 系统的本体模型，然后从维基的文章中提取实例，接着提取和描述 DBpedia 的概念和属性，以 RDF 转储和 SPARQL 端点访问中国既有的知识基地[14]，以期在 DBpedia 的 LOD 数据集（结构化知识的重要基础）中建设大规模的中文维基资源。

2.6　全球人类基因数据集集成和应用

随着基因组测序项目的快速增长，需要一套支持开放数据访问和协同工作的可视化系统和互动平台，特别是具有丰富功能和灵活配置框架的可定制的基因组浏览器[15]。北京大学生物信息中心开发的 ABrowse 是一个提供交互式浏览体验的基因组浏览器，可以进一步进行数据分析，以多个数据访问方法支持外部平台，并且终端用户可以创建用户空间，对资源进行存储、分享评论、注释和标注[16]。数据被存档和收集起来并且在适当时机发布是开放科学的重要举措，北京大学的这项开放数据实践，不仅提供了一个方便导入注释数据集的实用程序，而且所有源代码和技术文档都对外开放，供其他从事开放数据管理的团队参考。

2.7　全球医疗病例与特殊疾病治疗开发

心脏解剖和生理学的数学和统计模型在了解心脏疾病和治疗策略中发挥了至关重要的作用，而这类模型的准确性和预测能力，依赖于非侵入性成像数据集的广度和深度[17]。在医学领域中，已建立了心脏影像检查和相关的临床数据的大型数据库（Atlases）提供网络访问和开放数据共享，同时提供可视化参数描述，其所有软件开发的组件都是开源软件，根据 Mozilla 公共许可协议版本 1.1 免费提供[18]。任何科学结论都需要经过检验，运用这些结论从事诸如治疗等实践工作时，检验过程更为严格。基于这个逻辑，开放数据能够让同行快速取得相同的数据资料，从而加快检验过程和扩大可检验科学结果的专业人群。如果数据掌握在少数人手里而非开放数据，则这种科学检验的过程势必耗时费力，并且这些少数人还必须承担所有科学结论转化为实践应用的道德风险，基于此，开放数据会越来越多。

3　公众需求所产生的社会工具

在科技创新的需求下，许多开放数据工具纷纷出现，在海洋、地质、生物、药品、医疗等诸多领域目前已经有许多开放数据的应用工具值得关注和参考。然而，有别于理工农医，在社会科学中，特别是网络参与、政府治理和地理疆界等领域，也开始产生公众所需的社会工具，但这些社会工具尚未被充分开发。

3.1　全球网络终端用户的知识服务需求

人们希望根据自己的兴趣填补有限的宝贵时间，与此同时，文化传播机构也正努力吸引人们参与他们精心策划的文化活动，为此开发一个链接开放数据的 RDF/OWL 表示框架可以针对性地发送聚合事件，汇总、充实、提议这些事件。目前，从注册的所有用户群体中自动构造聚集全球用户配置文件（Profile – 型人）以推荐用户适当的事件信息，可以通过开放数据提供一个开放的、用户友好的系统平台[19]。这是因为开放数据能够丰富用户兴趣（End – User Interest）的条目，社群系统根据这些事件的项目分类，并通过智能索引以及网络上提供的链接开放数据集，增加用户可选性或者信息推荐的准确度。

3.2　政府治理技巧的开放数据需求

与科技创新的开放数据需求不同，科技创新主要考虑开放数据的预期受惠效果。而政府治理技巧的开放数据需求，则在于信息披露越多，承担政治责任的压力（政务官）越少，行政管理工作（事务官）的负责项目越少。欧洲议会和欧盟理事会的 2003/98/EC 启动的政府开放数据运动，是根据 W3C（World Wide Web Consortium）的数据描述建议提供政府数据的访问方法，这个项目包括开放数据的结构与格式、可以被重用的数据以及提供给公民和企业的新服务[20]等。目前，欧洲当局正在推动支撑数字经济和民主透明度的公共信息重用政策，例如西班牙的 Aporta 项目在公共管理部和工商贸易部的支持下，起草了“公共信息再利用法”[24]。这是因为公共部门信息的再利用是公开数据和开放政府中日益重要的组成成分。然而，与科技信息政策不同，公共信息政策的规划、制定、决策支撑工作等，需要首先考虑到公共行政及其他利益相关者是否达成一致理解，特别是对于开放数据的收集以及存储格式、发布数据方式等，这涉及到语义工具的管理方法和业务工作的使用目的[21]。然而一般教育体系毕业或者公共部门训练出来的主管并不具备行政经验、公共关系、信息技术和多种外语等多重能力，因此相关

的开放数据尚未被充分开发。

3.3 地理勘探界的开放数据战略

2010 年 4 月 1 日,英国地形测量局推出 OS Open Data 在线地图门户网站,允许用户浏览、下载或者开发简单的数据应用,人们可以访问英国的地理信息(Geographic Information, GI)并且提供相关应用,以促进政府透明度和鼓励更多地理信息数据注入[22]。更重要的是,它强化英国对某些特殊的地理位置(特别是边缘岛屿与海权领地等)的信息管治,使得国内外更加容易发现和访问[23]。关于英国地形测量局面向"公众"的地理标识和开放数据服务,值得进一步观察其他方面和各国反应。

4 开放数据所面临的政策问题

在"最大程度地利用数据"和"最大程度地保护安全隐私"之间,一个具体的挑战是发展电子信息创建和管理的创新和替代方法[25]。开放数据集有许多好处,然而隐私问题阻碍了建立开放健康数据。在美国,为确保公众使用的健康数据的安全机制,开放数据必须符合美国健康保险流通与责任法案(Health Insurance Portability and Accountability Act, HIPAA)的隐私规则的要求,开放数据的管理技术必须采用模拟攻击和匹配试验的风险识别,经过鉴定能够实现开放数据的建立原则[27]后才能实现。与科技创新息息相关的开放科学及开放数据管理,其技术手段和科技信息政策紧密联系、共同发展。

生物信息学领域的科研人员一直在推动开源软件和开放数据的发展,但是,隐私问题特别是个人基因组数据等,形成对重要数据集的访问限制,尤其凸显在基因组测序数据的大规模共享数据上。首先是基因组测序对象的基本辨识,即使有非常详细的个人特征,也有可能出现未预料到的基因型,在患者同意的时间之后,有可能释放更多个人医疗记录;其次,针对基因隐私问题的各种计算策略,可以采取一个切割格式化数据集的方式,使得部分共享同时确保个人基因隐私;再者,相比大公司和基因组研究中心,小型实验室直接面对个人隐私和数据安全,采用云计算可能对下载和计算大型数据集等进行更多控制[28]。然而,关键问题在于对数据管理存在法律问题和技术方法,例如"知情同意程序"的争议是当前保护隐私的措施在未来科技中可能失效,或者安全的云计算环境的标准规范是否在未来依然有效等问题。

在以数据为中心的"大生物学"学科,目前面临的三大挑战是:全面的数据标准、鼓励个别科学家共享数据、适当的基础设施和支持。因为关联开放数据的存在,所以克服技术的问题不大,但是对生命科学的异构数据的文化缺乏了解,使得单纯想以技术解决问题的作法,受到学科传统和现实环境的诸多无形障碍和干扰[29]。为了使研究数据得到充分利用,生物科学社群开始倡议技术和奖励机制以支持互操作性,促进开放科学与文化的增长[26]等。这些案例说明公共共享数据的框架描述,既要满足学科专业知识和科研人员的需求,又要熟悉掌握开放数据管理机制的技术和科技信息政策。

5 开放数据管理的预科学基础

强大的元数据方法和标准化发展可以提高数据的访问,但并不足够应付当前开放数据生态(地理、生命、社会科学数据集的合成规律)的发散性和异质性[30],需要一套良好的规划来解决诸如可执行的工作流程、数据重现性、所捕获的数据源、数据保存和复原、使用的归属和确认等一系列问题。

数据密集型的科学有处理庞大数据量的挑战,然而同样艰巨的挑战是跨学科数据的多样性,特别是研究数据,还需要数据集彼此的相互联系来理解复杂的系统性问题(如环境变化及其影响)。研究数据管理方法是面向复杂的跨学科问题的解决方式。虽然技术是处理数据密集型科学跨学科维度的关键因素,然而与过去那些分布式异构数据的不同之处在于,它需要更简单、更灵活、更有效的技术,更重要的是,有一个技术和文化适应的需求[31]。所以,积极发展战略性的科技数据管理的全谱段生态链有其必要性。

在面对开放的、关联的、实用的、安全的数据集合时,图书馆学对信息组织的原则和方法是面对众多学科的异同嵌入科研实际工作细节中,是累积成功经验最多和汲取失败教训最多的学科。这使我们有信心但也谨慎地提出一些短期阶段性强化服务战术和长期可持续性发展战略,以促进整体科学数据生态系统的健康化以及和谐社会经济技术的演进。

(致谢:感谢评委的评审和编辑部的校勘。)

参考文献：

[1] Garriga – Portola M. Open Data? Yes, But in a Sustainable Way [J]. *EI Profesional de la Informacion*, 2011, 20(3):298 – 303.

[2] Peset F, Ferrer – Sapena A, Subirats – Coll I. Linked Open Data and Open Data. Its Impact in the Field of Libraries and Information Science[J]. *EI Profesional de la Informacion*, 2011, 20(2):165 – 173.

[3] Granell C, Abargues C, Diaz L, et al. Interlinking Geoprocessing Services[C]. In: *Proceedings of the 2nd International Conference on Advanced Geographic Information Systems, Applications, and Services*. IEEE Computer Society, 2010:99 – 104.

[4] O'Boyle N M, Guha R, Willighagen E L, et al. Open Data, Open Source and Open Standards in Chemistry: The Blue Obelisk Five Years on[J]. *Journal of Cheminformatics*, 2011, 3: 37.

[5] Mir S, Staab S, Rojas I. An Unsupervised Approach for Acquiring Ontologies and RDF Data from Online Life Science Databases[C]. In: *Proceedings of the Semantic Web: Research and Application, 7th Extended Semantic Web Conference (ESWC2010)*. Berlin, Heidelberg: Springer – Verlag, 2010: 319 – 333.

[6] Torres R, Snoeij P, Geudtner D, et al. GMES Sentinel – 1 Mission [J]. *Remote Sensing of Environment*, 2012, 120:9 – 24.

[7] Margheriti L, Chiaraluc L, Voisin C, et al. Rapid Response Seismic Networks in Europe: Lessons Learnt from the L'Aquila Earthquake Emergency[J]. *Annals of Geophysics*, 2011, 54(4):392 – 399.

[8] Gower A C, Spira A, Lenburg M E. Discovering Biological Connections Between Experimental Conditions Based on Common Patterns of Differential Gene Expression [J]. *BMC Bioinformatics*, 2011, 12(7):381.

[9] Gaulton A, Bellis L J, Bento A P, et al. ChEMBL: A Large – scale Bioactivity Database for Drug Discovery[J]. *Nucleic ACIDS Research*, 2012, 40(D1):D1100 – D1107.

[10] The International Molecular Exchange Consortium. Submit Your Data [EB/OL]. [2011 – 10 – 27]. http://www.imexconsortium.org/submit – your – data.

[11] Kerrien S, Aranda B, Breuza L, et al. The IntAct Molecular Interaction Database in 2012[J]. *Nucleic Acids Research*, 2012, 40: D841 – D846.

[12] IntAct. Molecular Interaction Database[DB/OL]. [2012 – 03 – 08]. http://www.ebi.ac.uk/intact.

[13] Day N, Downing J, Adams S, et al. CrystalEye: Automated Aggregation, Semantification and Dissemination of the World's Open Crystallographic Data [J]. *Journal of Applied Crystallography*, 2012, 45(2):316 – 323.

[14] Wang Z C, Wang Z G, Li J Z, et al. Knowledge Extraction from Chinese Wiki Encyclopedias[J]. *Journal of Zhejiang University – Science C*, 2012, 13(4):268 – 280.

[15] Kong L, Wang J, Zhao S Q, et al. A Browse – A Customizable Next – generation Genome Browser Framework[J]. *BMC Bioinformatics*, 2012, 28(13):2.

[16] Center of Bioinformatic. A Map Like Rice Genome Browser[EB/OL]. [2012 – 03 – 08]. http://arabidopsis.cbi.edu.cn/.

[17] Fonseca C G, Backhaus M, Bluemke D A, et al. The Cardiac Atlas Project – An Imaging Database for Computational Modeling and Statistical Atlases of the Heart [J]. *Bioinformatics*, 2011, 27(16):2288 – 2295.

[18] Mozilla. Mozilla Public License[EB/OL]. [2011 – 10 – 22]. http://www.mozilla.org/MPL/MPL – 1.1.txt.

[19] Coppens S, Mannens E, Pessemier T D, et al. Unifying and Targeting Cultural Activities via Events Modelling and Profiling[J]. *Multimedia Tools and Applications*, 2012, 57(1):199 – 236.

[20] Ferrer – Sapena A, Peset F, Aleixandre – Benavent R. Access to and Reuse of Public Data: Open Data and Open Government[J]. *EI Profesional de la Informacion*, 2011, 20(3):260 – 269.

[21] Debruyne C, De Leenheer P, Spyns P, et al. Publishing Open Data and Services for the Flemish Research Information Space[C]. In: *Proceedings of the 30th International Conference on Advances in Conceptual Modeling: Recent Developments and New Directions*. Berlin, Heidelberg: Springer – Verlag, 2011:389 – 394.

[22] Lilley B. The Ordnance Survey Open Data Initiative[J]. *Cartographic Journal*, 2011, 48(3):179 – 182.

[23] McLaren R, Waters R. Governing Location Information in the UK [J]. *Cartographic Journal*, 2011, 48(3):172 – 178.

[24] Marcos – Martin C, Soriano – Maldonado S L. Reuse of Public Sector Information and Open Data in the Spanish and European Context. Aporta Project[J]. *EI Profesional de la Informacion*, 2011, 20(3):291 – 297.

[25] Pettenati M C, Pirri F, Giuli D. e – Profile Management as a Basic Horizontal Service for the Creation of Specialized e – Services[C]. In: *Proceedings of the 1st International Conference on Exploring Services Science*. 2010:259 – 263.

[26] Sansone S A, Rocca – Serra P, Field D, et al. Toward Interoperable Bioscience Data[J]. *Nature Genetics*, 2012, 44(2):121 – 126.

[27] Emam K E, Arbuckle L, Koru G, et al. De – identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset[J]. *Journal of Medical Internet Research*, 2012, 14(1):e33.

[28] Greenbaum D, Sboner A, Mu X J, et al. Genomics and Privacy: Implications of the New Reality of Closed Data for the Field[J]. *PLOS Computational Biology*, 2011, 7(12):e1002278.

[29] Thessen A E, Patterson D J. Data Issues in the Life Sciences[J]. *Zookeys*, 2011(150):15 – 51.

[30] Reichman O J, Jones M B, Schildhauer M P. Challenges and Opportunities of Open Data in Ecology [J]. *Science*, 2011, 331(6018):703 – 705.

[31] Parsons M A, Godoy O, LeDrew E, et al. A Conceptual Framework for Managing Very Diverse Data for Complex, Interdisciplinary Science [J]. *Journal of Information Science*, 2011, 37(6):555 – 569.

(作者 E – mail:gulp@mail.las.ac.cn)